NAVAL AEROSPACE MEDICAL RESEARCH LABORATORY
51 HOVEY ROAD, PENSACOLA, FL 32508-1046

NAMRL-1408

STATISTICAL ANALYSIS OF ISOPERFORMANCE ISSUES IN NAVY FLIGHT TRAINING

D. J. Blower

ABSTRACT

It is important to resolve the issue of whether extra information can help assign the probability for failure of a pilot or Naval Flight Officer (NFO) in some phase of flight training. This assigned probability for failure could be based simply on empirical data gathered over some relatively long period of time. However, if scores from selection tests, personality tests, vision exams, psychomotor tests, and the like could serve as cogent information about the probability for failure, then the probability for failure could be revised upwards or downwards based on an individual's standing on these variables. In addition, it would be interesting to find out if candidates could "trade off" high scores on one class of tests for low scores on a different class of tests, but still achieve the same level of performance. Here, level of performance is defined as the probability for failure. This brings us into contact with the idea of isoperformance. In this analysis, we examine two classes of predictor variables where candidates might trade off high scores for low scores, yet still achieve the same level of performance. The first class consists of cognitive information processing variables. Scores for the final academic grade from Aviation Pre-Flight Indoctrination (API) will serve as an example of this class. The second class consists of personality variables. We will use scores from the Pilot Biographical Inventory (PBI), a subcomponent of the Aviation Selection Test Battery, as a surrogate for scores on personality tests to be administered in future research on isoperformance.

Acknowledgments

I would like to acknowledge the continuing support of LT Henry P. Williams, Head of the Aviation Selection Division at NAMRL. Thanks also to Dr. Ken Ford, Director of the Institute for Human and Machine Cognition at the University of West Florida, for the support he has provided over the years.

INTRODUCTION

It is important to resolve the issue of whether extra information can help assign the probability for failure of a pilot or Naval Flight Officer (NFO) in some phase of flight training. The probability for failure could be based simply on empirical data gathered over some relatively long period of time. For example, such records indicate that about 25% of pilot trainees will fail to meet the required standards for success in Primary, Intermediate, or Advanced flight training after completing Aviation Pre-Flight Indoctrination (API). However, if scores from selection tests, personality tests, vision exams, psychomotor tests, and the like could serve as cogent information about the probability for failure, then numbers like the 25% predicted failures could be revised upwards or downwards based on an individual's test scores. Improved decisions about the career path for individuals could then be made on the basis of this extra information.

In addition, it would be interesting to find out if candidates could "trade off" high scores on one class of tests for low scores on a different class of tests, yet still achieve the same level of performance. Here level of performance is defined as the probability for failure. This brings us into contact with the idea of isoperformance. An article in the journal *Human Factors* by Jones and Kennedy (1) prompted our current interest in applying isoperformance to the analysis of selection and training data.

In this analysis, we examine two classes of predictor variables where candidates might trade off high scores for low scores, yet still achieve the same level of performance. The first class consists of cognitive information processing variables. Scores on the final academic grade from API will serve as an example of this class. The second class consists of personality variables. We will use scores from the Pilot Biographical Inventory (PBI), a subcomponent of the Aviation Selection Test Battery (ASTB), as a surrogate for scores on personality tests to be administered in future research on isoperformance.

As part of another project called the Pilot Prediction System (PPS), we have constructed a rather large and comprehensive data base consisting of various selection and training variables. A subset of this data base contains information on $N = 1,120$ Navy and Marine Corps candidates who entered pilot flight training from 1993 to the beginning of 1998. Crucial to the present analysis, this data base tells us if a student pilot failed some phase of flight training and, if so, which particular phase.

A previous paper (2) presented the quantitative rationale for the analysis carried out in this report. An information theoretic formula was derived based on Bayesian model evaluation to compute whether a tentative model could be accepted or rejected. The model assigned values to n cells of a contingency table. These values, labeled as Q_i, were the theoretical probabilities that a subject would fall into one of the n cells.

THE DATA BASE

The data analyzed in this report consist of a subset of the overall PPS data base. We selected 1,120 Navy and Marine Corps pilot candidates who were in training during the period from 1993 through January 1998. Scores on the various subtests of the ASTB and all the grades from the academic ground school (API) portion of training prior to actual flight training are part of this data base. We will concentrate on one of the subtests from the ASTB, the Pilot Biographical Inventory (PBI), and the final overall grade from API called the Navy Standard Score (NSS).

The raw score on the PBI is transformed into one of seven discrete categories so that PBI = $3, 4 \cdots 9$ with 3 being the lowest score and 9 the highest score. The API NSS is tranformed into one of six discrete categories, API = $1, 2 \cdots 6$ with, again, 1 representing low scores and 6 representing high scores from ground school. Thus, API represents one predictor variable from the class of cognitive information processing variables, and PBI represents one predictor variable from the class of personality variables. Table 1 shows the breakdown of the raw scores on the PBI and API variables into their respective discrete categories.

One criterion variable will be used in the subsequent analysis. This criterion variable simply records whether a candidate failed some later phase of flight training after API. This includes Primary, Intermediate, and Advanced

Table 1: PBI and API Raw Scores Broken Down Into their Respective Discrete Categories.

PBI Raw Score	PBI Category	API Raw Score	API Category
39–43	3	27–40	1
44–48	4	41–45	2
49–54	5	46–50	3
55–59	6	51–55	4
60–64	7	56–60	5
65–69	8	61–66	6
70–91	9	–	–

stages of flight training. A total of 281 candidates failed some phase of flight training in this data base, and a total of 839 candidates successfully completed all phases of training through advanced flight training yielding a failure rate of 25.1%.

Contingency Tables and the Raw Data

The data from the PBI, API, and criterion variables will be placed into contingency tables for analysis. Each contingency table consists of n cells with all $N = 1,120$ candidates placed into one, and only one, of the n cells. Each one of these n cells is defined as the intersection of one level of a predictor variable with one level of the criterion variable. The raw data consist of frequency counts inserted into each of the cells.

We will first examine each predictor variable separately for its impact on attrition. For example, consider the relationship between PBI and attrition. The table on the left-hand side of Fig. 1 shows the contingency table for the seven levels of the PBI variable and the two levels of the criterion variable. Each cell consists of the intersection of one of the predictor variable levels with a criterion variable level. There are thus $n = 7 \times 2 = 14$ cells in this first contingency table. The small number at the top of each cell indicates the cell number, while the larger number in each cell indicates the frequency count (the number of candidates) who fell into this particular intersection of categories.

		Fail	Pass	
	3	1 7	2 13	20
	4	3 40	4 82	122
	5	5 55	6 150	205
PBI	6	7 65	8 179	244
	7	9 49	10 160	209
	8	11 43	12 130	173
	9	13 22	14 125	147
		281	839	1,120

		Fail	Pass	
	1	1 52	2 44	96
	2	3 60	4 112	172
	3	5 70	6 229	299
API	4	7 72	8 265	337
	5	9 19	10 159	178
	6	11 8	12 30	38
		281	839	1,120

Figure 1: The raw data for analyzing the relationship between each separate predictor variable and attrition. The contingency table on the left shows the relationship between PBI and attrition while the contingency table on the right shows the relationship between API and attrition. The numbers in each cell are frequency counts.

For example, the fourth cell contains the intersection of PBI = 4 and PASS. There were 82 candidates from the total of 1,120 who fell into this intersection of categories, i.e., who scored a 4 on the PBI and passed all phases of flight training. The sum over the seven rows of the PBI categories must equal 281 for the FAIL column and must equal 839 for the PASS column.

The relationship between API scores and attrition is shown in a similar contingency table on the right-hand side of Fig. 1. This table differs from the one just discussed only in the fact that now $n = 12$ because the table is displayed with six rows and two columns reflecting the division of the API variable into six discrete categories.

To address the issue of isoperformance, we will need to analyze the two predictor variables together with the criterion variable. We can conveniently display the n cells for this situation as two contingency tables with separate tables for PASS and FAIL. Figure 2 shows the frequency count data for two 7 × 6 tables. These are the data that will be analyzed in a later section on the issue of isoperformance. Both tables show the intersection of each level of the PBI variable with each level of the API variable for one of the two levels of the criterion variable. Here $n = 84$ cells. The 21st cell, for example, shows that 16 candidates who failed some phase of flight training fell into the cell defined by the intersection of having a PBI score of 6, and an API score of 3.

Fail

API

PBI	1	2	3	4	5	6	
3	1 3	2 1	3 1	4 1	5 0	6 1	7
4	7 10	8 13	9 7	10 4	11 4	12 2	40
5	13 12	14 7	15 15	16 17	17 3	18 1	55
6	19 11	20 11	21 16	22 20	23 5	24 2	65
7	25 9	26 10	27 11	28 15	29 4	30 0	49
8	31 7	32 11	33 13	34 9	35 2	36 1	43
9	37 0	38 7	39 7	40 6	41 1	42 1	22
	52	60	70	72	19	8	281

Pass

API

PBI	1	2	3	4	5	6	
3	43 1	44 1	45 2	46 6	47 3	48 0	13
4	49 6	50 8	51 21	52 30	53 13	54 4	82
5	55 5	56 24	57 43	58 45	59 28	60 5	150
6	61 10	62 25	63 49	64 51	65 39	66 5	179
7	67 7	68 25	69 41	70 50	71 31	72 6	160
8	73 9	74 15	75 36	76 39	77 27	78 4	130
9	79 6	80 14	81 37	82 44	83 18	84 6	125
	44	112	229	265	159	30	839

Figure 2: To study the issue of isoperformance, we need at least two predictor variables and their relationship with attrition. To display such a relationship, we use two contingency tables, one for each of the two levels of the criterion variable.

Notice that the marginal totals running along the right-hand side of each contingency table in Fig. 2 match the entries in the left-hand table of Fig. 1. The marginal totals running along the bottom of each table match the corresponding entries in the right-hand table of Fig. 1.

MODELS AND PROBABILITIES

The frequency counts shown in the contingency tables of Figs. 1 and 2 are assumed to reflect an underlying probability attached to each of the n cells. These probabilities are labeled as $Q_1, Q_2, \cdots Q_i, \cdots Q_n$. For convenience, we will call this set of probabilities for the n cells the Q_i. Some assigned probabilities are compatible with the observed frequencies, but most probabilities that could be assigned to the Q_i are incompatible with the raw data. We would like to determine which of these assignments can be rejected by the data and which are supported by the data. Any proposed assignment of probabilities to the Q_i will be called a *model.*

In the companion paper (2), a formula was derived from the perspective of Bayesian model evaluation. The number computed by this formula is compared to a χ^2 distribution with an appropriate number of degrees of freedom. If the number falls into the upper 5% region of the χ^2 distribution, the proposed model is rejected; conversely, if the number falls into the lower 95% region of the distribution, the assigned probabilities are accepted as being compatible with the data.

We will now illustrate this formula for the data shown in the two contingency tables of Fig. 1. This exercise is in preparation for the more significant analysis of isoperformance data in Fig. 2. Equation (9) from Reference [2] is rewritten here as Equation (1):

$$2N\sum_{i=1}^{n} f_i \ln\left(\frac{f_i}{Q_i}\right) \sim \chi^2\left(\nu\ df\right) \qquad (1)$$

where N is the total number of candidates in the contingency tables. In this particular analysis $N = 1,120$. f_i stands for the relative frequency in the ith cell so that $f_i = N_i/N$ where N_i is the actual frequency count in the ith cell. Q_i is the probability assigned to the ith cell by any given model.

Let us first use Equation (1) to assess the impact of PBI scores on attrition. We will initially suggest two simple models for the Q_i, but both of these models will be rejected. Then we will find a third model that will be accepted. These models are labeled respectively as models $\mathcal{M}_A, \mathcal{M}_B$ and $\mathcal{M}_C$.

Simple explanations should always be checked before more complicated models are proposed. With that in mind, $\mathcal{M}_A$ is a model that says that all 14 Q_i for the 14 cells of the PBI contingency table are equal. In other words, the assigned probability for failing is equal to the probability for passing, and, furthermore, there is no relationship between PBI scores and attrition. Under this model each $Q_i = .0714$.

Equation (1) returns a value of 577.32 for this model. We compare this to a χ^2 distribution with $\nu = 13$ *df*. The degrees of freedom are calculated by considering n and subtracting the number of constraints on the Q_i. Since there is only one constraint on the Q_i, namely, that they must sum to 1, $\nu = n - 1 = 13$. A value of 22.36 marks the dividing point between the upper 5% region and the lower 95% region of the χ^2 distribution with 13 *df*. A value of 577.32 obviously falls into the upper 5% region of this χ^2 distribution. Therefore, this model is rejected and we must search for another model that the data can support.

Model $\mathcal{M}_B$ takes notice of the prior existing knowledge that about 25% of all candidates fail one of the later stages of flight training after API. The seven Q_i that together make up the probability of failure will sum to .25 and the remaining seven Q_i that together make up the probability of passing will sum to .75. However, we still retain the hypothesis that PBI scores have no impact on attrition. To mirror this hypothesis, the probabilities are made equal within the PASS and FAIL levels of the criterion variable. This is a slightly more complicated model than considered in model $\mathcal{M}_A$. We are inserting information only about the relative probabilities of passing and failing and nothing else. See Table 2 for a detailed list of the Q_i values for model $\mathcal{M}_B$. The sum under the two *Cell Probability* columns is rounded up to two decimal places.

Equation (1) returns a value of 286.49 for $\mathcal{M}_B$. Since one extra constraint has been introduced, the appropriate degrees of freedom for the χ^2 distribution is now $\nu = n - 2 = 12$. A value of 21.03 marks the dividing point between the upper 5% region and the lower 95% region of the χ^2 distribution with 12 *df*. Because a value of 286.49, while a marked improvement over Model $\mathcal{M}_A$, still falls into the upper 5% region of the χ^2 distribution, $\mathcal{M}_B$ must also be rejected.

Because the simple explanations of Models $\mathcal{M}_A$ and $\mathcal{M}_B$ have been clearly rejected, we are now allowed to hypothesize some minimal structure to our next class of models. We will test whether there is a relationship between PBI scores and attrition in the following model for the Q_i. The following model is just one example from the class of models that the data support.

Table 2: Specification of the Fourteen Q_i Values for Model $\mathcal{M}_B$.

Row	Q_i Fail	Cell Probability	Q_i Pass	Cell Probability
1	Q_1	.0357	Q_2	.1071
2	Q_3	.0357	Q_4	.1071
3	Q_5	.0357	Q_6	.1071
4	Q_7	.0357	Q_8	.1071
5	Q_9	.0357	Q_{10}	.1071
6	Q_{11}	.0357	Q_{12}	.1071
7	Q_{13}	.0357	Q_{14}	.1071
		.2500		.7500

We will hypothesize that PBI scores *are* related to the probability of failure. We expect that low PBI scores are associated with an elevated probability of failure while high PBI scores indicate an increased chance of success. Furthermore, we postulate that there should be some reasonably smooth function that relates probability of failure and the PBI scores. That is, the probability of failure should decrease for each unit increase in the PBI score. We retain the desirable feature from Model $\mathcal{M}_B$ that the Q_i that make up the probability of failure, irrespective of anything else, should add up to .25.

As an example of constructing such a model, refer to Table 3. The last column shows the desired characteristics for the probability for failure, that is, that it be a monotonic, decreasing function of PBI. These values in the last column are calculated from Bayes's Formula.

Table 3: Model $\mathcal{M}_C$ Which Hypothesizes Some Structure to the Q_i Such That the Probability for Failure is a Decreasing Monotonic Function of PBI Scores.

PBI	Q_i Fail	Cell Probability	Q_i Pass	Cell Probability	Probability Fail
3	Q_1	.01	Q_2	.01	.50
4	Q_3	.04	Q_4	.08	.33
5	Q_5	.05	Q_6	.13	.28
6	Q_7	.06	Q_8	.16	.27
7	Q_9	.05	Q_{10}	.15	.25
8	Q_{11}	.03	Q_{12}	.11	.21
9	Q_{13}	.01	Q_{14}	.11	.08
	Total	.25		.75	

The following calculation shows how Bayes's Formula is used to find the probability of failure given that a candidate had a PBI score of 3.

$$\begin{aligned} P(\text{Fail}|\text{PBI} = 3) &= \frac{P(\text{Fail and PBI} = 3)}{P(\text{Fail and PBI} = 3) + P(\text{Pass and PBI} = 3)} \\ &= \frac{Q_1}{Q_1 + Q_2} \end{aligned}$$

$$= \frac{.01}{.01 + .01}$$

$$= .50$$

The value returned by Equation (1) for Model $\mathcal{M}_C$ is 15.73. We added another constraint on the Q_i in order to arrive at $\mathcal{M}_C$, so the critical value demarcating the upper and lower regions of the χ^2 distribution for 11 *df* is 19.68. 15.73 falls into the lower 95% region where Model $\mathcal{M}_C$ can be accepted.

To lend some sense of the variation in the probability of failure inherent in the class of good models, look at Table 4. The probabilities assigned to the Q_i shown in the table above represent another good model, call it Model $\mathcal{M}_D$. The value from Equation (1) for this model is 16.44 and thus it also falls into the acceptable region of the χ^2 distribution. Instead of the marked functional relationship between lower PBI scores and higher probability for failure evident in model $\mathcal{M}_C$, no extra information is contained in these PBI scores. The probability for failure does not change with changing PBI score and could have been predicted solely from the historical knowledge of the probability of failure.

Table 4: Another Acceptable Model, call it Model $\mathcal{M}_D$, Which Shows a Different, Much Less Dramatic Function of Probability for Failure Given PBI Scores.

PBI	Q_i Fail	Cell Probability	Q_i Pass	Cell Probability	Probability Fail
3	Q_1	.0043	Q_2	.0130	.25
4	Q_3	.0250	Q_4	.0740	.25
5	Q_5	.0450	Q_6	.1330	.25
6	Q_7	.0540	Q_8	.1580	.25
7	Q_9	.0470	Q_{10}	.1410	.25
8	Q_{11}	.0390	Q_{12}	.1160	.25
9	Q_{13}	.0357	Q_{14}	.1150	.24
		.2500		.7500	

We could find, in similar fashion, a large number of acceptable models. Some of these would show an interesting functional relationship between probability of failure and PBI scores just like Model $\mathcal{M}_C$, and some would show more of the flat-line relationship exhibited in Model $\mathcal{M}_D$. The point is that there is some uncertainty about the exact form of the functional relationship between probability of failure and PBI score. To express this uncertainty, we should attach some error bars to the point estimates of the probability of failure at each discrete PBI score. More generally, we think of a confidence band attached to the curve expressing the functional relationship between probability of failure and PBI scores. The confidence band will be larger at either end of the PBI range and smaller in the middle because the frequency counts are larger in the middle and smaller at the ends.

Figure 3 provides a rough, qualitative sense of what these error bars would look like for the PBI scores. The circles indicate a good point estimate for the probability of failure at each PBI score by being placed at the actual relative frequency as given by the data. The width of the error bars represent the extremes of the acceptable models.

A more precise analysis would find the average of all the acceptable models and, at the same time, calculate the standard deviation around the average. This kind of analysis would provide the best functional curve and its associated confidence band. Such calculations are the concern of *Bayesian Model Averaging* (3), but concentrating on this topic would take us too far afield for the present investigation. We are content, at this point, merely to locate some good models at the extremes of acceptability to obtain some general sense of the uncertainty of the functional relationship.

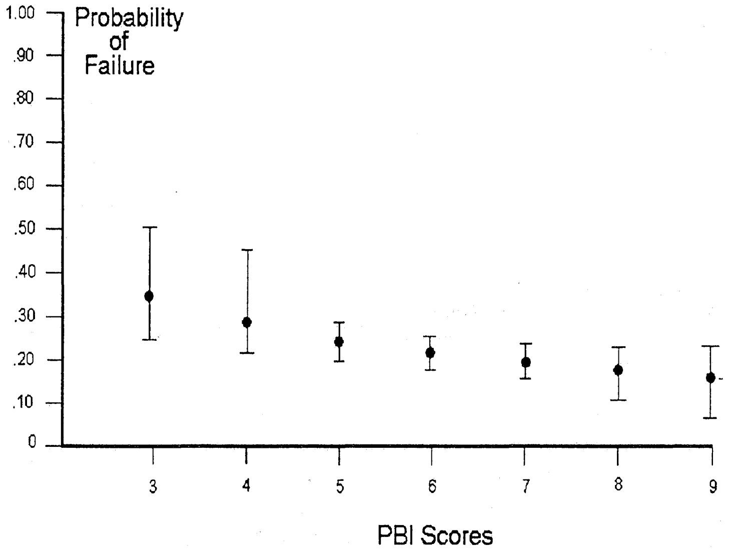

Figure 3: A curve relating PBI scores to probability of failure. The error bars attached to each point estimate give some idea of the range of acceptable models for the given data.

Another good model, $\mathcal{M}_E$, which is intermediate between Models $\mathcal{M}_C$ and $\mathcal{M}_D$, is shown in Table 5. There is a functional relationship between probability for failure and PBI score, but it is not as strong as $\mathcal{M}_C$. However, it is not quite the flat-line relationship of $\mathcal{M}_D$ either. The χ^2 value for $\mathcal{M}_E$ is 6.39, which falls into acceptance region for 11 *df*.

Table 5: Another Acceptable Model, $\mathcal{M}_E$, Which Shows an Intermediate Relationship between Probability for Failure and PBI Scores.

PBI	Q_i Fail	Cell Probability	Q_i Pass	Cell Probability	Probability Fail
3	Q_1	.008	Q_2	.015	.35
4	Q_3	.035	Q_4	.080	.30
5	Q_5	.050	Q_6	.135	.27
6	Q_7	.057	Q_8	.160	.26
7	Q_9	.045	Q_{10}	.135	.25
8	Q_{11}	.040	Q_{12}	.125	.24
9	Q_{13}	.015	Q_{14}	.100	.13
	Total	.250		.750	

Figure 4 shows three curves dictated by the models $\mathcal{M}_C$, $\mathcal{M}_D$, and $\mathcal{M}_E$. This gives an example of the variation in functional relationships to be expected from the class of good models.

The same kind of analysis just described for the PBI scores as a single predictor variable can be performed on the API scores. Model $\mathcal{M}_F$ where all 12 Q_i are equal to .08333, meaning that the probability for failure is equal to 50% for all API scores, is rejected with a value of 748.90. The critical value for the χ^2 distribution that reflects only this universal constraint is 19.68. Because this contingency table has only $n = 12$ cells, $\nu = n - 1 = 11$ *df*.

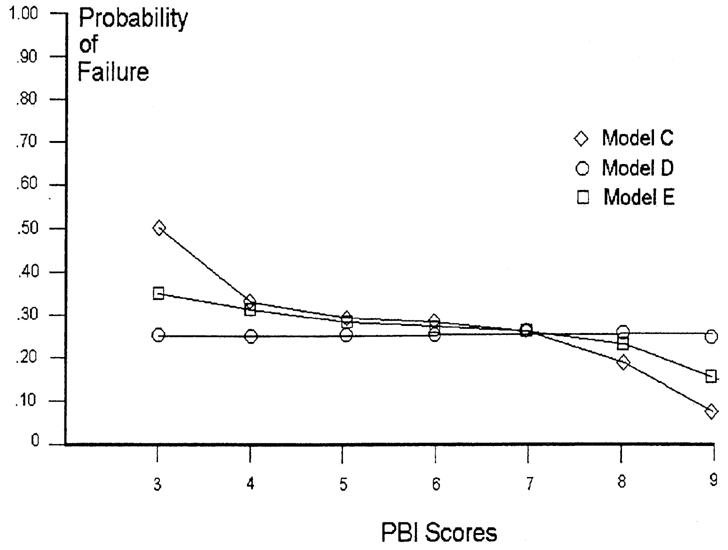

Figure 4: Three representative curves showing functional relationship between probability for failure and PBI scores.

Model $\mathcal{M}_G$, where the six fail Q_i are each equal to .041667 and the six pass Q_i are each equal to .125 so that overall probability for failure equals .25 and probability for passing is equal to .75, is also rejected with a value of 458.01. The degrees of freedom drop to 10 because of the extra constraint on the Q_i and the critical χ^2 value is 18.31.

We search for another model that can be accepted by Equation (1). Model $\mathcal{M}_H$, as outlined in Table 6, is one such acceptable model. This model has a value of 9.84 and clearly fits into the lower 95% region of a χ^2 distribution with 9 *df* where the critical value is 16.92. Here we see that there is a stronger functional relationship between API scores and probability for failure than existed for PBI scores.

Table 6: Model $\mathcal{M}_H$ Hypothesizes a Certain Structure to the Q_i Such That the Assigned Probability for Failure is a Decreasing Function of API Scores.

API	Q_i Fail	Cell Probability	Q_i Pass	Cell Probability	Probability Fail
1	Q_1	.052	Q_2	.035	.60
2	Q_3	.059	Q_4	.090	.40
3	Q_5	.062	Q_6	.200	.24
4	Q_7	.058	Q_8	.250	.19
5	Q_9	.016	Q_{10}	.145	.10
6	Q_{11}	.003	Q_{12}	.030	.09
Total		.250		.750	

Just as for PBI scores, there are many models for assigning probability values to the Q_i that are acceptable. Table 7 presents another acceptable model, Model $\mathcal{M}_I$. This model shows a less pronounced influence of API scores on probability of failure. As before, we now present a model in Table 8, $\mathcal{M}_J$, that is intermediate between $\mathcal{M}_H$ and $\mathcal{M}_I$. It has a χ^2 value of 8.89 for 9 *df*.

Table 7: Model $\mathcal{M}_I$ Shows a Much Less Pronounced Impact of API Scores on Probability for Failure Than Model $\mathcal{M}_H$.

API	Q_i Fail	Cell Probability	Q_i Pass	Cell Probability	Probability Fail
1	Q_1	.041	Q_2	.044	.48
2	Q_3	.050	Q_4	.120	.29
3	Q_5	.064	Q_6	.200	.24
4	Q_7	.066	Q_8	.240	.22
5	Q_9	.024	Q_{10}	.120	.17
6	Q_{11}	.005	Q_{12}	.026	.16
	Total	.250		.750	

Table 8: Model $\mathcal{M}_J$ Shows an Intermediate Impact of API Scores on Probability for Failure As Compared to Models $\mathcal{M}_H$ and $\mathcal{M}_I$.

API	Q_i Fail	Cell Probability	Q_i Pass	Cell Probability	Probability Fail
1	Q_1	.045	Q_2	.040	.53
2	Q_3	.054	Q_4	.100	.35
3	Q_5	.063	Q_6	.204	.24
4	Q_7	.062	Q_8	.237	.21
5	Q_9	.020	Q_{10}	.128	.14
6	Q_{11}	.006	Q_{12}	.041	.13
	Total	.250		.750	

Figure 5 is similar to Fig. 3 in that it shows error bars attached to each point estimate of the assigned probability for failure at each of the six API scores. These error bars are just rough indications of where some extreme, but still acceptable, models could be located. Models $\mathcal{M}_H, \mathcal{M}_I$, and $\mathcal{M}_J$ are all examples of acceptable models that lend a sense of the uncertainty in the functional relationship between API scores and the probability for failure. Figure 6 plots the three curves represented by Models $\mathcal{M}_H, \mathcal{M}_I$, and $\mathcal{M}_J$.

ISOPERFORMANCE ISSUES

Having become somewhat familiar with the techniques surrounding a single predictor variable and one criterion variable, we can now transition to the case of two predictor variables and one criterion variable. This is the simplest case where we can talk about isoperformance and is the main focus of the current investigation.

When we employ the concept of isoperformance, we seek that combination of two or more predictor variables that result in the same probability for failure. In the particular application treated here, we are interested in trading off good scores on one class of variables with bad scores on another class of variables such that overall performance (assigned probability for failure) remains the same. For example, can good scores on API compensate for low scores on the PBI, or vice versa?

Fortunately, there is nothing new demanded from a theoretical point of view to address this issue. Bayesian model evaluation remains the framework within which the quantitative analysis is carried out. Equation (1) is still applicable, and we continue to exploit the idea of assigning probabilities to the n cells of the contingency tables.

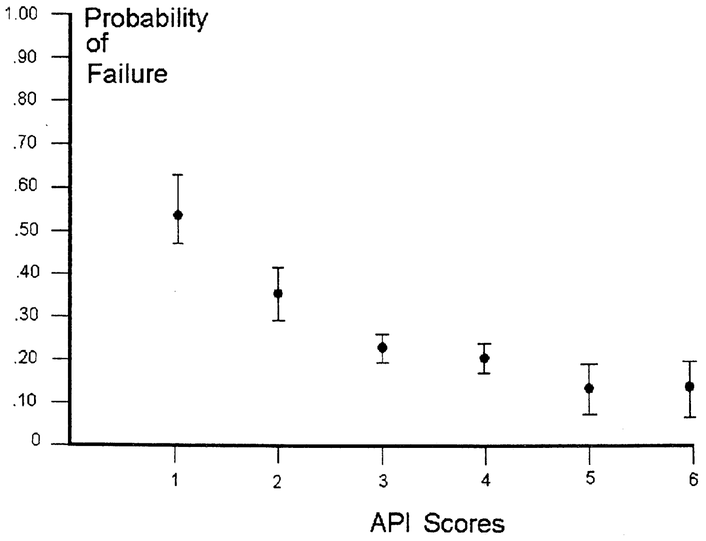

Figure 5: Error bars attached to the point estimates of the functional relationship between API scores and probability for failure.

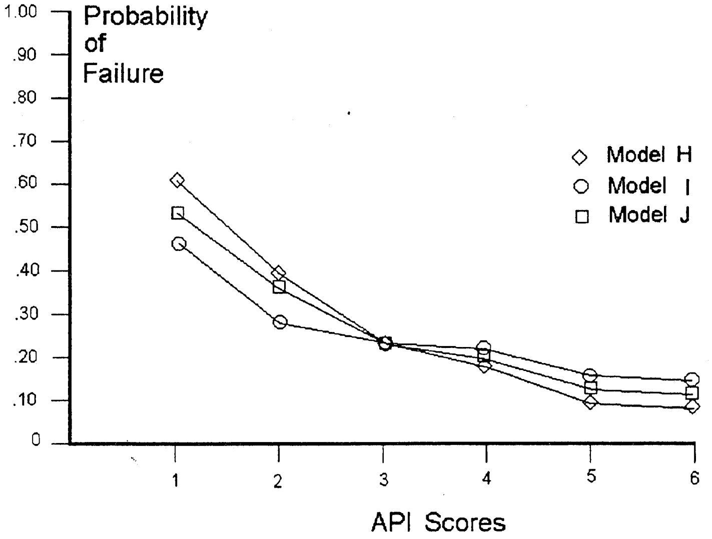

Figure 6: Three curves illustrating the range of functional relationships captured by some acceptable models.

Granted, the n cells of the contingency tables become more numerous with each addition of a new predictor variable. Compare the tables of Fig. 1 where $n = 14$ or 12 with the isoperformance table of Fig. 2 where $n = 84$.

One additional difference is that we now have two pieces of data, the scores from both the PBI and the API, as given information in Bayes's Formula. Suppose that we have as specific pieces of information a score of 5 on the PBI and a score of 3 on the API. Bayes's Formula yields

$$P(\text{Fail}|\text{PBI} = 5 \text{ and API} = 3) = \frac{P(\text{Fail and PBI} = 5 \text{ and API} = 3)}{P(\text{Fail and PBI} = 5 \text{ and API} = 3) + P(\text{Pass and PBI} = 5 \text{ and API} = 3)} \tag{2}$$

Refer back to Fig. 2 to find the Q_i values to plug into this equation:

$$P(\text{Fail}|\text{PBI} = 5 \text{ and API} = 3) = \frac{Q_{15}}{Q_{15} + Q_{57}} \tag{3}$$

If Q_{15} were assigned the value of .01 and Q_{57} the value of .03, then the probability for failing conditioned on this information would equal .25.

The goal is still to find good models that represent an acceptable fit to the observed frequency data and to reject other models because of the large discrepancy when matched with the data. The measure of such a discrepancy is provided by Equation (1) when the value it returns is placed into a region of high or low probability of a χ^2 distribution. The class of good models should be explored to provide some sense of the uncertainty to be attached to the hypothesized functional relationship between the probability for failure and the two predictor variables of interest.

We begin by showing, just as we did in the one predictor case, that certain simple benchmark models that posit no relationship between the predictor variables and the criterion variable can be rejected. Next, we present some examples of models that do fit the data, but which also represent minimum insertion of extra structure to explain the proposed relationship. This is done with explicit recognition of the precepts of Occam's razor.

Start with a model, $\mathcal{M}_K$, that includes the two predictor variables and, which by analogy to Models $\mathcal{M}_A$ and $\mathcal{M}_G$, says that all the Q_i are equal. This implies that probability for failing conditioned on the two predictor variables is always equal to .50. This model is clearly rejected, as is the next model, $\mathcal{M}_L$, analogous to Models $\mathcal{M}_B$ and $\mathcal{M}_H$. These models assigned equal Q_i to all fail cells and equal Q_i to all pass cells, but set the overall probability for failing at .25 and the overall probability for passing at .75. These models imply that the probability for failing, given any PBI and API score, remains constant at .25 and that, therefore these two predictor variables provide no useful information beyond what is already known from the historical data.

The first acceptable model is Model $\mathcal{M}_M$ shown in Table 9. The entries in this table are the probability for failing given the PBI score as indexed by the row and the API score as indexed by the column. All these values were calculated according to Bayes's Formula as illustrated above in Equations (2) and (3).

This model was found by assigning Q_i values close to the observed frequencies. The combination of low PBI and API scores seems to raise the probability for failing (or lower the probability for passing). Intermediate scores hover around the .25 level. A combination of high scores on both variables offers some evidence that probability for passing is increased (or the probability for failing is lowered). The important thing here is that, contrary to models $\mathcal{M}_K$ and $\mathcal{M}_L$, there appears to be some sort of functional relationship between the information in the predictor variables and the probability assigned for failing.

Just as before with the one predictor case, there are many acceptable models for the two predictor case that exhibit a good fit of the assigned Q_i to the observed frequencies. Model $\mathcal{M}_M$, however, has the defect of not providing a smooth declining relationship between the discrete levels of the PBI and API scores and probability for failing.

For example, within the API = 1 level the probability for failing starts out at .74, decreases to .64, then jumps

Table 9: One Acceptable Model, $\mathcal{M}_M$, for the Two Predictor Variables Used in the Analysis of Isoperformance.

	API					
PBI	1	2	3	4	5	6
3	.74	.53	.31	.23	.23	.33
4	.64	.50	.24	.16	.20	.25
5	.71	.22	.25	.27	.14	.17
6	.53	.31	.24	.28	.13	.26
7	.53	.29	.20	.22	.15	.16
8	.43	.39	.27	.19	.07	.19
9	.38	.33	.15	.12	.05	.13

$\chi^2 = 19.64$, $df = 81$, critical value=102.65

back up to .71, then back down again to .53. Within the PBI = 4 level the probability for failing declines steadily through API categories 1–4, but then starts to increase for categories 5 and 6.

Table 10 presents another acceptable model, $\mathcal{M}_N$, which exhibits a somewhat better behavior in this regard. The probability for failing shows a smoother functional relationship than Model $\mathcal{M}_M$. This is bought at the price of a larger χ^2 value, in other words, the assigned Q_i had to deviate from the observed frequencies a little bit more than Model $\mathcal{M}_M$ to provide this kind of relationship.

Table 10: Another Acceptable Model, $\mathcal{M}_N$, for the Two Predictor Variables Which Exhibits a Smoother Functional Relationship between Increasing Scores and Declining Probability for Failing.

	API					
PBI	1	2	3	4	5	6
3	.90	.53	.44	.30	.29	.29
4	.75	.50	.25	.16	.14	.13
5	.71	.30	.25	.27	.14	.11
6	.53	.29	.24	.28	.13	.11
7	.53	.28	.20	.22	.15	.16
8	.43	.28	.20	.19	.07	.08
9	.38	.29	.15	.12	.05	.07

$\chi^2 = 39.32$, $df = 81$, critical value=102.65

This second acceptable model provides the first glimpse of the variability associated with the functional relationship between the two predictor variables and probability for failing. We are already aware from our analysis of the single predictor variable that the class of good models provides a range of values at each probability. The isoperformance issues involved with two predictor variables must cope with the same problem. In fact, the variability is worse with two predictor variables because the frequencies at each intersection of the contingency table are smaller than the one predictor case.

As a third acceptable model to shed some light on the variability problem, consider Model $\mathcal{M}_O$ in Table 11. As an example, concentrate on the entry in the first row and first column from Tables 10 and 11. We see that

Table 11: Another Acceptable Model, $\mathcal{M}_O$, for the Two Predictor Variables, Which Together with Models $\mathcal{M}_M$ and $\mathcal{M}_N$ Give Some Idea of the Variability to be Expected in the Functional Relationship.

	API					
PBI	1	2	3	4	5	6
3	.69	.47	.41	.24	.29	.29
4	.57	.50	.25	.16	.25	.27
5	.50	.30	.26	.27	.17	.30
6	.50	.29	.24	.28	.15	.20
7	.50	.28	.20	.22	.15	.27
8	.47	.33	.24	.19	.11	.18
9	.44	.29	.15	.13	.13	.29

$\chi^2 = 40.35$, $df = 81$, critical value=102.65

P(Fail|PBI = 3 and API = 1) can range at least from a high value near .90 to a low value near .69. Likewise, in the last row and last column, P(Fail|PBI = 9 and API = 6) can range at least from .29 to .07.

A fourth and final model, $\mathcal{M}_P$, is presented in Table 12 because it affords the opportunity to make some interesting remarks about isoperformance. Model $\mathcal{M}_P$ was constructed such that as many entries as possible were equal to .25. This model is just on the wrong side of acceptability with a χ^2 value of 106.76 as compared to the critical value of 102.65 for 81 df. Nonetheless, with minor tweaking of the .25 values, it could have been made acceptable, but keeping the .25 values simplifies matters and illustrates the point so nicely that we retain Model $\mathcal{M}_P$.

Table 12: A Fourth Model, $\mathcal{M}_P$, Where the Concept of Isoperformance Can Be Easily Illustrated.

	API					
PBI	1	2	3	4	5	6
3	.71	.33	.25	.25	.25	.25
4	.64	.25	.25	.25	.25	.25
5	.64	.25	.25	.25	.25	.17
6	.53	.25	.25	.25	.25	.17
7	.52	.25	.25	.25	.15	.15
8	.38	.25	.25	.18	.14	.12
9	.27	.25	.15	.12	.10	.09

$\chi^2 = 106.76$, $df = 81$, critical value=102.65

All those entries in Table 12 with a value of .25 represent occasions where flight students can trade off a score on the first predictor variable with a score on the second predictor variable, yet still achieve the same level of performance. For example, a low PBI score of 3 if combined with an API score of 3, 4, 5, or 6 will still result in an assigned probability for failing of .25. A low API score of 2 can be offset by any PBI score greater than 3. An API score of 3 and a PBI score of 7, or an API score of 4 and a PBI score of 6 represent intermediate scores that can be traded off quite freely for the same probability for failing.

SUMMARY

Ascertaining the probability for success of a student over all phases of flight training is a prerequisite for optimizing decisions relevant to naval aviation selection and training. We have presented here a quantitative technique for assigning probability for success based on observed frequency counts. The technique is solidly based on an approach derived from Bayesian model evaluation and has been shown to connect with information theoretic concepts and the χ^2 distribution.

Specifically, one would like to know what happens to the probability of success as a function of information about scores on selection test batteries, personality tests, vision exams, psychomotor skills, race, gender, college major, and so on. The essential question boils down to: *Does this information change the probability for success from what was known without these scores?*

As a natural outgrowth to the solution of this problem, we are able to address issues in isoperformance. That is, we are interested in the question of whether a subject can trade off low scores on one set of skills like cognitive information processing with high scores on another set of skills like motivation to become a pilot. Based on the analysis here, cognitive information processing variables like the API scores and a personality test surrogate like the PBI scores both significantly alter the probability for success from its base line value of 75%. These results were examples of the simplest application to a single predictor variable.

The next issue is the consideration of the joint effect of both API scores and PBI scores. Here again, there were significant changes in the probability for success as a function of knowledge about two predictor variables. Obtaining the lowest scores on both the API and PBI materially lowered the probability of success, while high scores on both these variables raised the probability of success from its base line value.

Critical to the isoperformance question, we could also determine which combination of scores resulted in a trade off where the probability of success remained essentially the same. For example, a relatively low API score of 2 could be traded off for any PBI score of say 6 or higher in order to remain at a 75% probability of success.

An important adjunct to this kind of analysis is a determination of the variability involved in making a point estimate for the probability of success as a function of the predictor variables. This was accomplished in an illustrative manner by sampling from the class of all good models. Such samples lend a rough idea of the variability that must be attached to any one functional relationship that might be presented as an explanation for the probability of success. A more rigorous analysis would have sampled more extensively from the class of good models and then formed an average of the predictions from each of these good models as weighted by the posterior probability of each model. Such an analysis is planned for the future with an even larger data set.

One feature that a frequency count clearly reveals is that the sample size fluctuates markedly over the various discrete levels of the predictor variables. There are fewer counts at the extreme ends of the predictor variables and many more at the middle levels. The situation becomes progressively worse when tables for two or more predictor variables in combination with a criterion variable are assembled. There may only be a few students or even no students at the intersection of low probability cells. This is quite understandable if we think that the predictor variable scores are roughly normally distributed. We are then asking for the combination of scores falling, for example, into the lower 2% of the distribution of the first variable with the highest 2% of the second variable. This naturally means that the estimation at this combination of extreme scores will have a much higher attached variability.

For example, there would be very few cases in any data base for API scores of 1 and PBI scores of 9. In fact, there were no cases in the data base analyzed in this report with this combination of scores together with the criterion measure of failing some phase of flight training. The consequence of this fact is that the attached variability will be much higher for point estimates of the probability of success given these scores as compared to equivalent estimates of probability of success for scores like API = 4 and PBI = 6 where many cases exist. This is unfortunate for isoperformance because it is exactly at these extreme combination of scores where most interest lies.

REFERENCES

1. Jones, M.B. and Kennedy, R.S. Isoperformance Curves in Applied Psychology. *Human Factors,* 38(1), 167–182, 1996.

2. Blower, D. J. Some General Quantitative Considerations for the Analysis of Isoperformance Curves. *NAMRL-1406,* Naval Aerospace Medical Research Laboratory, Pensacola, FL, October 1999.

3. Raftery, A.E. Bayesian Model Selection in Social Research (with Discussion). In P.V. Marsden (Ed.) *Sociological Methodology 1995,* pp. 111-163, Blackwell Publishers, Cambridge MA, 1995.

Reviewed and approved 19 JAN 2000

R. R. STANNY, Ph.D.
Technical Director

UNITED STATES NAVY
NAVAL AEROSPACE MEDICAL RESEARCH LABORATORY

This research was sponsored by the Office of Naval Research under work unit 62233N.0330.126-7801.

The views expressed in this article are those of the author and do not necessarily reflect the official policy or position of the Department of the Navy, Department of Defense, nor the U.S. Government.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY *(Leave blank)*	**2. REPORT DATE** 19 JAN 2000	**3. REPORT TYPE AND DATES COVERED**

4. TITLE AND SUBTITLE
Statistical Analysis of Isoperformance Issues in Navy Flight Training

5. FUNDING NUMBERS
62233N.0330.126-7801

6. AUTHOR(S)
David J. Blower

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Naval Aerospace Medical Research Laboratory
51 Hovey Road
Pensacola Fl 32508-1046

8. PERFORMING ORGANIZATION REPORT NUMBER
NAMRL-1408

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Office of Naval Research
800 North Quincy Street
Arlington, VA 22217-5660

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited.

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

It is important to resolve the issue of whether extra information can help assign the probability for failure of a pilot or Naval Flight Officer (NFO) in some phase of flight training. This assigned probability for failure could be based simply on empirical data gathered over some relatively long period of time. However, if scores from selection tests, personality tests, vision exams, psychomotor tests, and the like could serve as cogent information about the probability for failure, then the probability for failure could be revised upwards or downwards based on an individual's standing on these variables. In addition, it would be interesting to find out if candidates could ``trade off" high scores on one class of tests for low scores on a different class of tests, but still achieve the same level of performance. Here, level of performance is defined as the probability for failure. This brings us into contact with the idea of isoperformance. In this analysis, we examine two classes of predictor variables where candidates might trade off high scores for low scores, yet still achieve the same level of performance. The first class consists of cognitive information processing variables. Scores for the final academic grade from Aviation Pre-Flight Indoctrination (API) will serve as an example of this class. The second class consists of personality variables. We will use scores from the Pilot Biographical Inventory (PBI), a subcomponent of the Aviation Selection Test Battery, as a surrogate for scores on personality tests to be administered in future research on isoperformance.

14. SUBJECT TERMS
Military selection, Contingency Tables, Choice of models, Bayesian statistics

15. NUMBER OF PAGES
26

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT UNCLASSIFIED	**18. SECURITY CLASSIFICATION OF THIS PAGE** UNCLASSIFIED	**19. SECURITY CLASSIFICATION OF ABSTRACT** UNCLASSIFIED	**20. LIMITATION OF ABSTRACT** SAR